文字通りの感情で伝わるボイスチャットシステム

宮岡 拓也 * 青木 秀憲 † 宮下 芳明 *

概要. 音声発話をするとき，人は口調によって発話内容にニュアンスを加えている．この装飾のニュアンスと発話意味が同じ感情を表現しているとは限らない．そのため，聞き手に発話内容が正確に伝わらず，誤解してしまう可能性がある．それを防ぐためには，口調によって装飾される方向が，発話内容と揃っていることが必要と考えられる．本稿では，発話内容を正しく理解するための支援として，文章を解析してネガティブかポジティブかを判定した上で，そのニュアンスを音声変換して発話音声に付与する仕組みを提案する．このシステムにより，発話内容と聞き手が受ける印象が一致し，発話内容を正しく理解することへの補助を可能にする．

1 はじめに

音声によるコミュニケーションは情報の伝達における主要な手段の 1 つである．人は発話する際，口調を変化させ，話の内容にさらなる情報を音声に加えている．このとき，装飾の意図と発話内容は必ずしも一致するとは限らない．また，言語情報と発話者の感情が矛盾した場合，聞き手が受け取る情報のうち，言語情報は 7%と報告されている [1]．そのため，発話者の伝えたい内容が聞き手に正しく伝わらない場面があると考えられる．本研究では，聞き手が発話内容を正確に理解するための支援として，発話内容に含まれる文章を解析してネガティブかポジティブかを判定した上で，そのニュアンスを音声変換して発話音声に付与する仕組みを提案する．このシステムにより，ネガティブな発話内容は実際の口調よりもネガティブに，ポジティブな発話内容は実際の口調よりもポジティブに伝わるようになる．これは特に，発話内容よりも口調を重視してとらえがちな聞き手が発話内容を正しく理解することを助け，コミュニケーション支援につながると筆者らは考えている．

2 関連研究

音声に関する遠隔コミュニケーションを支援する研究は，コミュニケーションにポジティブな影響を与えることを目的としているものが多い．Costa らは参加者のトーンを変化させることにより対立する状況下におけるコミュニケーションの不安の解消に貢献することを明らかにした [2]．Wang らは振幅と周波数を制御することにより説得力を変化させられることを報告した [3]. 二瓶らは参加者が任意のタイミングで表情と音声のピッチを肯定的に変容させるシステムを使用することによって，遠隔コミュニケーションが活発化することを示した [4]．これらは音声全体に対し，一時的に音声加工を行っている．一方，音声の語尾に着目した研究では，西原らはオンライン会議において語尾のピッチをリアルタイムで上昇させるシステムを用いることで，聞き手が話し手に対し元気である印象を抱く傾向が増加したことを示した [5].

音声加工を利用したコミュニケーション支援では，文章の意味を考慮せず，文章全体または一部をそのまま加工している．本研究では，文章の意味を強調させて相手に伝え，聞き手が内容を正確に把握することを促進させる．

3 ピッチシフト機能に関する予備実験

3.1 実験方法

ピッチシフトにより，聞き手にネガティブ・ポジティブが伝わるか予備実験を行った．3 つの文章でピッチを変更させたサンプルとピッチを変更していないサンプルを男声・女声それぞれ用意した．音声は JVS コーパス [6] に含まれる文章に対し，感情分析機能によってニュートラルと判断された以下の 3 種類を使用した.

- 時間領域と、空間領域で共通する処理手法は、フィルタリングによる、入力信号の強化である。
- 南西部ウォーレンは、ベイアーマンファームズと、フィッツジェラルドの地区で、構成される。
- 肝臓への酸素供給は、肝動脈と、低圧系の門脈を介して、行われている。

これらの音声を実験参加者に聴かせ，音声がポジティブに聴こえるか回答させた．回答は 5 段階とし，

* 明治大学
† ソニーグループ株式会社

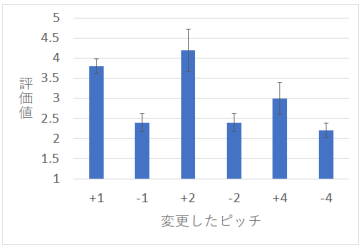

図 1. 変更したピッチとポジティブさ

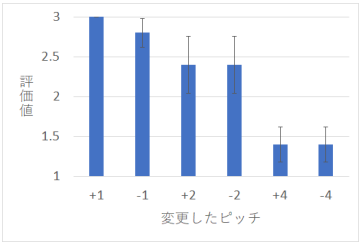

図 2. 変更したピッチと不自然さ

1＝「ネガティブ」，5＝「ポジティブ」とした．また，音声が不自然に聴こえるか回答させた．回答は 3 段階とし，1＝「不自然」，3＝「自然」とした．実験参加者は 30〜50 代の男性 5 名であり，実験では文章や話者の順番をランダムに変更して行った．

3.2 実験結果および考察

結果を図 1 および図 2 に示す．ピッチの値は半音単位，記号はピッチの上下，評価値は実験参加者が回答した値の平均値である．ピッチを半音 2 個分上げた音声が一番ポジティブ，ピッチを半音 4 個分下げた音声が一番ネガティブに感じられた．また，ピッチを半音 2 個分上げた音声はあまり不自然に感じなかった一方，ピッチを半音 4 個分下げた音声は不自然に感じられた．音声が不自然に変化するとコミュニケーションに悪影響が出る恐れがあるため，実際のシステムで採用する場合，ネガティブな音声はピッチを半音下げたものを使用することが好ましいと考えられる．

4 システム概要

本稿で提案するシステムについて述べる．システムの概略図を図 3, 聞き手側のインタフェース画面を図 4 に示す．本システムは大きく分けて以下の 4 つの機能を持つ．

- 発話した音声をテキストに変換する音声認識機能
- テキストからテキストそのものが持つ感情を分析する機能
- ボイスメッセージと感情分析の結果を保存するサーバ

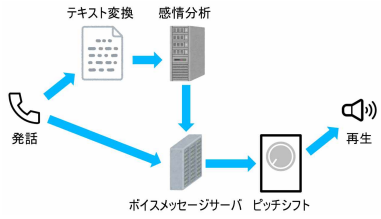

図 3. 提案システムの概略図

図 4. 聞き手側のインタフェース画面

- 感情分析の結果を元に音声にピッチシフトを行う機能

4.1 音声認識機能

本システムで利用する音声認識機能は Web ブラウザ上で利用可能な Web Speech API[7] で提供されている SpeechRecognition を使用する．発話した内容はすべて API を通じてテキスト情報に変換され，それを感情分析するサーバに送信する．

4.2 感情分析

感情分析では BERT[8] ベースの感情推定モデルを利用し，テキスト情報を受け取るとネガティブかポジティブかを判定するプログラムを実行する．音声認識機能から受けとったテキスト情報は適宜，感情分析を行い，その文章がネガティブかポジティブかを判定し，ボイスメッセージサーバに結果を渡す．

4.3 ピッチシフト機能

ボイスメッセージサーバから取得した情報に合わせて前章の予備実験で得たパラメータを用いてピッチシフトを行い，その音声データをボイスメッセージサーバに送信する．これにより，聞き手側はピッチシフトをかけられたデータを取得することになる．

5 おわりに

本稿では，聞き手が文字通りの感情で理解するためのシステムを提案した．今回はネガティブ・ポジティブを判定したものを提案したが，今後は感情をより細かく分類し，その感情に応じた音声加工を行えるようにしたいと考えている．

参考文献

[1] Mehrabian, Albert. Communication without words, Psychological Today, Vol.2, pp.53-55, 1968

[2] Costa, Jean. Jung, Malte F. Czerwinski, Mary. Guimbretière, François.Le, Trinh. Choudhury, Tanzeem. Regulating Feelings During Interpersonal Conflicts by Changing Voice Self-Perception, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp.1-13, 2018

[3] Wang, Tzu-Yang. Kawaguchi, Ikkaku. Kuzuoka, Hideaki. Otsuki, Mai. Effect of Manipulated Amplitude.Frequency of Human Voice on Dominance. Persuasiveness in Audio Conferences, Vol.2, No.CSCW, 2018

[4] 二瓶 芙巳雄, 田口 和佳奈, 中野 有紀子, 深澤 伸一, 赤津 裕子. 多人数遠隔コミュニケーションにおける肯定的感情表出支援の効果と支援適用タイミングの決定, 情報処理学会論文誌, Vol.62, No.2, pp.761-771, 2021

[5] 西原 宗太郎, 渡邉 拓貴, 寺田 努, 塚本 昌彦. オンライン会議において相手に与える印象を変化させるためのリアルタイム語尾ピッチ変換システム, マルチメディア，分散協調とモバイルシンポジウム 2021 論文集, Vol.31 pp.188-196, 2021

[6] Takamichi Shinnosuke. Mitsui Kentaro. Saito Yuki. Koriyama Tomoki. Tanji Naoko. Saruwatari Hiroshi. JVS corpus: free Japanese multi-speaker voice corpus, 2019

[7] Web Speech API - Web APIs | MDN `https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API` (閲覧日：2021 年 11 月 18 日)

[8] Jacob, Devlin. Ming-Wei, Chang. Kenton, Lee. Kristina, Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019

未来ビジョン

コミュニケーションにおいて感情は重大な要素であり，音声はコミュニケーションで用いられる手段の 1 つである．2020 年以降，新型コロナウイルスの流行により，マスクを着用したコミュニケーションが身近に行われるようになり，表情から感情を読み取ることが難しくなった．また，ビデオ会議やボイスチャットなどオンライン上でのコミュニケーションも普及したが，カメラを使用した会話は心理的負担やプライバシー等の問題があり，積極的に使わない人も存在する．そのため，コミュニケーションにおける音声の重要性はこの数年で増加していると考えられる．音声から感情を読み取ることがより求められるようになった一方，会話する際に感情が表立って出ない人や発話内容と感情が一致していない人もいる．そのような人の感情を読み取ることや推定することは難しく，コミュニケーションに支障をきたす原因となる．また，発話者自身が意識的に感情を表出することは精神的負担を強いる．このような状況を解消するためには，言語情報と感情を一致させることで，聞き手が発話内容を正しく理解できるにすることが重要であると推測される．

本研究は，言語情報と感情を一致させることで，コミュニケーションにおける齟齬を防止し，聞き手が発話内容を正しく理解することへの支援を目指しており，本稿はその第一歩である．本稿で提案したシステムは，発話内容がネガティブかポジティブかに応じてピッチシフトを行い，内容と口調を一致させて伝えるものである．今後は，様々な感情に応じた音声加工をすることで，感情をより正確に伝え，コミュニケーションをスムーズに行なったり，発話内容が聞き手により正しく伝わるよう支援していきたいと考えている．